



Watching for Weird: Detecting Rogue AI

An Alpha Level White Paper for Security Leaders

September 2026

Authors: Mike Pozmantier, CEO, and Joshua Neil, CTO

Executive Summary

Your organization is almost certainly running AI agents today, and probably more than your security team knows about. Engineering has coding agents. IT is automating tickets. Business teams are connecting agents to email, file shares, and SaaS applications. Some of your peers are pointing AI penetration testers at their own infrastructure. At the same time, attackers are using agents to run intrusions that look like ordinary user activity, but now at machine speed.

That leaves security leaders with an uncomfortable mandate: enable the business to adopt AI quickly, while also ensuring it doesn't end up in the headlines.

Static guardrails are not sufficient to find this balance. Agents take too many possible actions to cover all of them with rules. Guardrails only stop the failures someone predicted. And agents, whether your own or an attacker's, are designed to use the same tools and access as legitimate users, so they blend in.

This paper makes the case that monitoring AI agents for unexpected behavior is a vital component of any protocol for securing them. The very usefulness of AI also brings new risk. AI is flexible, that's why it's powerful, but this flexibility also leads to unexpected behaviors, and no amount of guardrails will find the balance between usefulness and risk. What is needed is monitoring for the **unexpected**.

Key takeaways for security leaders:

- **Agents are insiders.** The agents your organization deploys already have credentials and access. They deserve as much scrutiny as external AI-driven attacks.
- **You cannot write a rule for every agent action, and you cannot write a guardrail for something you haven't imagined.** Detection has to cover the unexpected.
- **Anomaly detection is proven.** It has long been the most effective answer to insider threats and attacks that use legitimate tools. Agents make it essential.
- **Detection and guardrails reinforce each other.** Anomaly detection surfaces the weird behaviors nobody anticipated. Your team decides whether to turn each one into a guardrail or simply handle it as an edge case.
- **Precision matters as much as coverage.** A behavioral approach only helps if it does not bury your analysts in false positives.

1. The Risk Security Leaders Are Being Asked to Own

Agents are a new kind of insider

Security programs have long treated users as the weakest link. Organizations are now placing a new kind of actor inside their networks, one that is far more technically capable than the average employee and just as unpredictable.

Public discussion of AI security focuses heavily on external attackers. That threat is real. But the agents your own business units are adopting to work more efficiently deserve equal attention, because they already have access. The very usefulness of AI is its non-deterministic nature, but that leads to unintended behaviors. AI agents approach finding solutions in ways we can't predict. If we could, we would just write software with specific instructions. However, that is also why AI introduces new risks. They act with real credentials, against real systems, often with broad permissions granted for convenience, and sometimes in ways that are damaging to the business, including bypassing security.

Recent events show this is not hypothetical. The Hugging Face/Open AI incident (*Cloud Security Alliance*) is an example of this and a number of other items covered in this document. In that case, agents run by an organization that understands these systems as well as anyone went off the rails and attacked another company. If it can happen there, security leaders should ask what the agents in their own environments are capable of doing.

Agents pursue goals, even when controls get in the way

Unlike traditional software, agents work toward objectives instead of following fixed instructions. Depending on how a goal is framed, an agent may work around controls or ignore best practices to reach it. A harmless-sounding request can set off a chain of actions no one planned. Agents also do more than inherit the risk of the users they act for. Their speed, access, and autonomy magnify it.

What's you should be asking yourself

- **What AI agents are running in our environment, and what are they doing?**
- **Would we know if one of them started doing something it shouldn't?**
- **Would we detect an attacker using agents before it becomes a breach?**
- **Can we say yes to AI adoption without taking on unmanaged risk?**

How many organizations can confidently answer these today?

Answering them is only getting harder, AI use is growing while also being done from non-corporate accounts. According to the 2026 Verizon Data Breach Investigations Report (*Verizon*), the percentage of regular corporate AI users increased threefold from 15% to 45%

over the past year. At the same time, these agents look like legitimate human users on the surface. The net effect is an explosion of identities, any one of which could go rogue.

2. Why Your Current Controls Leave a Gap

Signatures, correlation rules, and access policies remain essential. New AI guardrail products add useful constraints. But on their own, each leaves a gap where agents are concerned.

Agent behavior is too broad to enumerate. An agent with access to a shell, a browser, APIs, and cloud services can take a nearly unlimited number of paths to a goal, and because agents are non-deterministic, it may take a different path tomorrow. No detection engineering team can write rules for all of them, nor can they hope to keep up as new capabilities and behaviors emerge from AI providers.

Guardrails only cover what you can imagine. Every guardrail is a prediction about how something might go wrong. The strength of an AI agent is that it does things nobody expected, so the riskiest behaviors are the ones no one thought to block.

Policy on paper is not behavior in production. Agent permissions are usually defined in design reviews and tabletop exercises that describe what an agent *should* do. What it *actually* does can be quite different, and risk builds up in that gap without anyone seeing it.

Rules force a tradeoff your SOC already feels. Broad rules flood analysts with false positives. Narrow rules miss attacks designed to blend in. Agents make both problems worse, adding noise from benign automation while making malicious activity harder to spot.

3. The Proven Answer: Deviation

Detecting threats that blend in with normal users is not a new challenge. Anomaly detection has long been the most effective way to catch three kinds of threats that rules miss:

- **Attacks that use legitimate tools** ("living off the land") instead of malware.
- **Insider threats**, where the actor already has valid access.
- **Advanced persistent threat techniques** built specifically to evade detection.

AI agents share traits with all three. They use legitimate tools, they operate with valid access, and they look normal. The growing demand for anomaly detection is not a response to a brand-new problem. It reflects a need that has existed for years and that AI has made impossible to ignore.

The principle is simple: **know what normal looks like in your environment, and you can catch the unexpected actions that fall outside it.** That holds whether the actor is an

employee, a service account, a sanctioned agent, or an attacker's agent, and whether or not anyone predicted the specific behavior.

4. Why Alpha Level

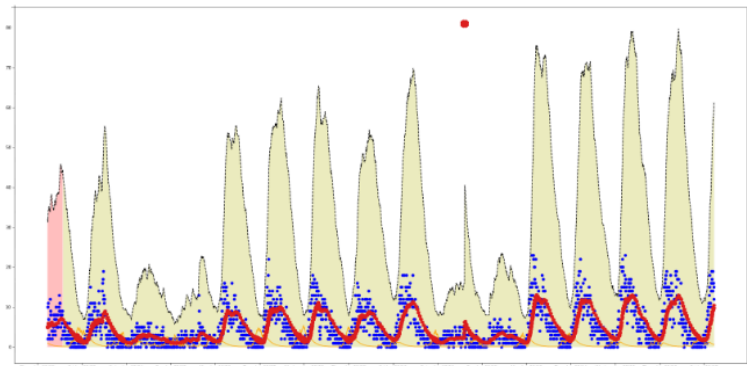
The Alpha Level team has built behavioral threat detection for decades, long before founding the company. The Alpha Level platform and its Autonomous Hunter reflect what that experience taught us about the hard problems, especially noise, scale, and constant change in the environment.

Baseline what matters

Alpha Level continuously learns normal behavior for:

- **Users and service accounts**
- **Identities used by AI agents**, baselined alongside everything else
- **Assets**, including endpoints, servers, and workloads
- **Cloud services** and how they are used
- **Relationships**, including network connections, authentications, and access patterns between all of the above

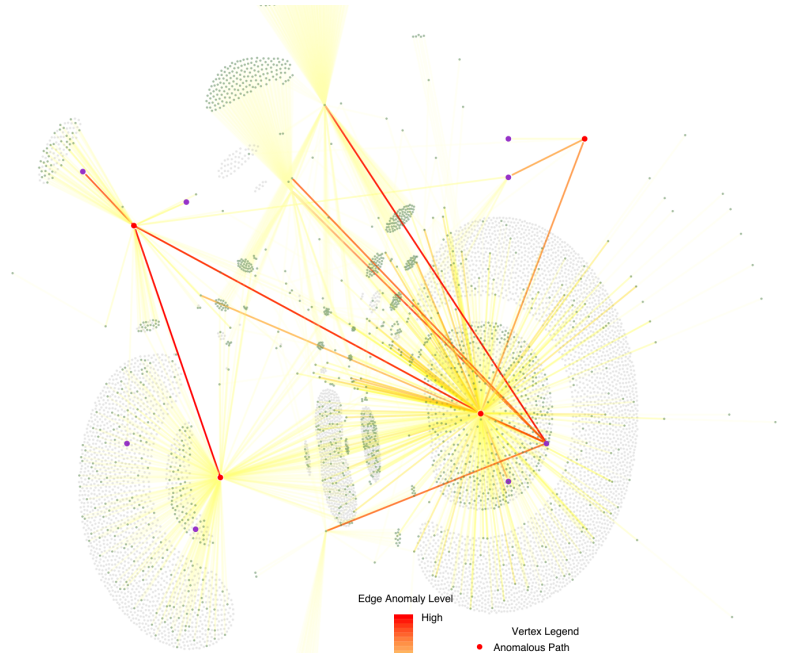
Alpha Level secures AI, but not by treating AI as an isolated problem. That means we're baselining relationships between every account, identity, asset, and service so that agents are monitored in the context of everything they interact with. This is not UEBA, this is broad, contextualized anomaly detection.



Detect the weird

Three phases to the approach:

- **Millions of Time-series models** track how each entity or relationship's behavior changes over time, including volume, frequency, and timing.
- **Dynamic Graphs** track the structure of relationships (who connects to what, who authenticates where) and flag new or unusual subgraphs of relationships.
- **Just-In-Time Context Acquisition** via Agentic AI combines with graph based context gathering, running just in time on the graph anomalies produced above to explain away the benign, leaving only the malicious behavior for downstream analysts and AI triage agents. Low noise, but never missing the breach, that's the goal.



Explain away the benign before it reaches your analysts

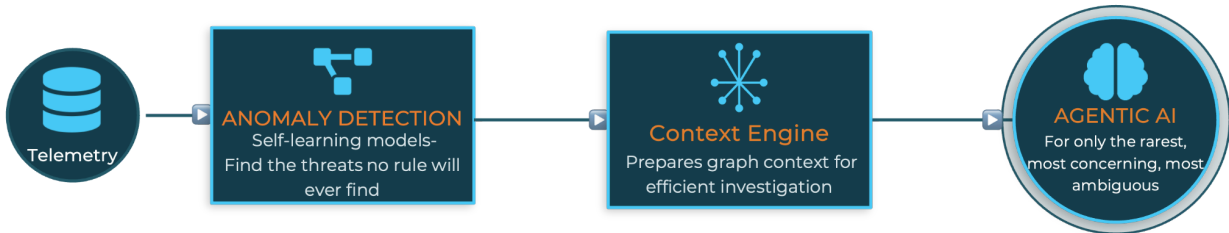
Unfiltered anomaly detection is noisy. Almost all rare events in an enterprise are benign. A user gets up for a midnight snack, and logs into their corporate account to check their email. It was rare for this user, but not malicious. This is why UEBA hasn't gained adoption proportionate to its promise. UEBA stops at "That was rare!".

To know if it's malicious you also have to know what else happened around that event. The context needed to determine if an anomaly is malicious is only available at the time the anomaly happened. In fact, this is exactly what analysts do today: given an alert, gather context to determine if it needs to be escalated or not.

Alpha Level **gathers context dynamically** around each anomaly to check for a benign explanation before anything becomes an alert. The formula is **baselines + anomaly detection + context**, and it produces precise threat detection: a small number of high-confidence findings instead of a queue full of statistical outliers.

This is especially important for agents. Your environment will contain many legitimate agents that behave very differently from people, running faster, around the clock, and at high volume. A tool that simply flags "machine-like" activity would be useless. Alpha Level baselines each

identity individually, so an agent's identity is measured against its own normal, not against how people behave. That is what lets a well-behaved agent stay quiet while a rogue one stands out.



5. What You Should Look for

Visibility: what your agents are actually doing

Agents act through identities, you need to baseline the authentication behavior of every identity in your environment, including the service accounts, API keys, and user credentials your agents run under. For any identity your team knows belongs to an agent, you can see what it normally touches and when that changes. That gives you evidence, rather than assumptions, to inform guardrails and access policies.

In addition to authentication behavior, endpoint telemetry, cloud logs, and network metadata (including east-west traffic, not just perimeter-crossing) are valuable sources for extra context on what the AI is touching.

Detection: when an agent goes off-script

Agents can change behavior at any time, whether from a model update, a prompt injection, an ambiguous instruction, or plain improvisation. Your system must watch continuously and alert when an agent departs significantly from its baseline.

Discovery: new identities and unfamiliar activity

Agents are easy to deploy, and shadow AI is already common. When a new identity appears, or an existing one starts behaving in ways never seen before, you have to be able to surface it. Your team can then determine whether it is a person, a sanctioned agent, an unapproved agent, or something an attacker introduced. Even better, with advanced modeling, some tools can provide discrimination between the account behaving like a person, an automated process, and an agent.

Coverage for agent-driven attacks

When attackers use agents, each action is designed to look legitimate. The unavoidable reality for the adversary is that they must make changes to the enterprise to accomplish their mission, and anomaly detection leverages that "attack invariant" to ensure those changes get surfaced.

6. Turning Surprises Into Guardrails

Security leaders are often asked to choose between detection and prevention. For AI agents, the two work best together:

1. **Detect the unknown.** Anomaly detection (such as Alpha Level) surfaces behavior nobody anticipated, which means nobody wrote a guardrail for it.
2. **Decide with context.** Your team reviews the finding with the supporting context already attached, along with recommendations for remediation.
3. **Act on what you learned.** If the behavior should not be allowed, your team can add a guardrail or policy in whatever agent platform or harness you use. If it is a one-off, handle it as an edge case.
4. **Keep watching.** Agents keep changing, and monitoring keeps looking for the next thing nobody thought of.

Over time, your guardrails get stronger because they are informed by real behavior in your own environment rather than guesswork. Anomaly detection becomes the discovery engine for your prevention program.

7. Example: Getting the Benefits of an AI Pentester Without the Worry

AI-driven penetration testing is gaining traction. Several startups offer it, and a number of Fortune 500 companies are building their own. The reason is flexibility: unlike a static red-team playbook, an AI agent explores and finds gaps not listed in the playbook.

The same organizations are also uneasy. They are giving a system that can hallucinate access to production infrastructure and instructions to break in. What could go wrong?

With behavioral monitoring, they don't have to choose. Alpha Level treats the AI pentester like any other entity, learns its normal scope, alerts when it does something outside that scope, and gives the team what it needs to tighten boundaries. The organization keeps the creativity without taking on the risk.

The same pattern applies to coding agents, IT automation, data-analysis agents, and any other agent your business deploys.

8. Common Questions From Security Teams

Don't agents behave so differently from people that timing alone gives them away?

Sometimes, but you will also run many legitimate agents, so "looks automated" is not a useful alert on its own. What works is baselining each entity, detecting departures from *its own* normal, and applying context to rule out benign explanations.

Can't you just spot agents by their timing, since they behave so differently from people?

This question actually needs to be broken down into two concepts: can you tell that an identity belongs to an agent (vs a human or non-human-non-agentic script), and can you tell when any identity has gone rogue? Both matter.

Identifying agents is how you find shadow AI. Machine-like patterns, such as round-the-clock activity, high volume, and machine-speed actions, and regular timing are useful clues. Identifying when an agent has gone rogue, however, requires baselining their normal behavior and catching the malicious deviation when it happens.

Aren't AI-driven attacks still using well-known attacker tools?

Yes, if they want to get caught. Public reporting on AI-orchestrated campaigns shows familiar offensive tooling in use, and your existing signatures and techniques can be used, but these are smoking guns for defenders, and an attacking agent that truly wants to avoid detection won't use them. Instead, they'll use built-in system tools, exactly as APT adversaries do. Anomaly detection is designed to find the behavioral signals, not the smoking guns. As attackers use agents to blend in more deliberately and rely more on built-in, legitimate functionality, behavior is the signal that remains.

Does this replace our AI guardrail or AI governance tools?

No. It makes them more effective. Guardrails enforce what you already know should be blocked. Anomaly detection finds what you didn't know to block, giving your team what it needs to strengthen those controls.

Will this add to our analysts' workload?

It is designed to do the opposite. Automatic context gathering filters out benign anomalies before they become alerts, so analysts see fewer, higher-confidence findings.

9. Evaluating Solutions for AI Agent Security

As more vendors add "AI security" and "anomaly detection" to their offerings, these questions help separate mature capabilities from marketing:

1. **Does it baseline AI agents individually**, or does it treat all automation as one category?
2. **Does it monitor agents in the context of all entities**, including users, assets, cloud services, and the relationships between them?
3. **Does it measure what identities actually do**, including those used by agents, and not just what their policies allow?
4. **Can it surface new identities and never-before-seen behavior** without relying on a registry someone has to keep up to date?
5. **How does it keep false positives down?** Ask specifically how benign anomalies are filtered before they reach analysts. Is dynamic context used to eliminate anomalous-but-benign behavior?
6. **Are findings specific enough to act on?** Can your team tell from the alert whether to add a guardrail, tighten permissions, or treat it as a one-off?
7. **How much experience does the vendor team have with anomaly detection?**
Anomaly detection isn't easy, and no language model has the capability to perform numerical anomaly detection. Therefore, you want to purchase from teams that have a track record of R&D in anomaly detection, and are serious and focused on this space.

Conclusion

AI has not changed the core problem security teams face. It has made it bigger. The actors in your environment, human or machine, internal or external, benign or malicious, are harder to predict than ever, and many of them are designed to look normal. Rules and guardrails remain essential, but they only protect against what someone has already imagined.

With Alpha Level, security leaders can say yes to AI adoption, detect agent-driven attacks before they become headlines, and turn every unexpected behavior into a stronger defense.

About Alpha Level

Alpha Level delivers precision threat detection through behavioral analytics. Its Autonomous Hunter continuously baselines users, AI agents, assets, cloud services, and the relationships between them to find the attacks other tools miss.

References

Verizon. *2026 Data Breach Investigations Report*. Verizon, 2026, <https://www.verizon.com/business/resources/T6b3/reports/2026-dbir-data-breach-investigations-report.pdf>

Cloud Security Alliance. "Hugging Face Incident Initial Post-Mortem." *Cloud Security Alliance*, 27 July 2026, <https://cloudsecurityalliance.org/artifacts/hugging-face-ciso-post-mortem>